

## Factors Characterizing En Route Operational Errors: Do They Tell Us Anything About Situation Awareness?

Francis T. Durso<sup>1</sup>, Todd R. Truitt<sup>1</sup>, Carla A. Hackworth<sup>1</sup>,  
Daryl Ohrt<sup>1</sup>, Janis, M. Hamic<sup>1</sup>, Jerry M. Crutchfield<sup>1</sup>,  
and Carol A. Manning<sup>2</sup>

<sup>1</sup> University of Oklahoma

<sup>2</sup> F.A.A. Civil Aeromedical Institute

It is the responsibility of the Air Traffic Control Specialist to maintain adequate separation of aircraft, as prescribed by the Air Traffic Control (ATC) Handbook (7110.65J). To perform satisfactorily, controllers must maintain a sufficient level of situation awareness (SA). Poor SA on the part of the controller may result in the loss of separation between two or more aircraft, otherwise known as an operational error. Each time an operational error occurs a lengthy, detailed investigation and report of the incident is completed by the facility concerned.

Operational error reports have previously been analyzed by a number of researchers (e.g., Rodgers & Nye, 1993; Schroeder & Nye, 1993; Redding, 1992). We took a similar approach to analyzing operational errors that occurred during 1993. We analyzed the 412 nonoceanic operational errors reported by air route traffic control centers (specifically en route centers) during calendar year 1993. From these reports, 388 were complete along the variables we considered and formed the database for the current study. An effort was made to determine those factors, if any, that characterized those errors. In particular, we attempted to determine the factors that influenced severity of the operational error. In each case, we used a stepwise multiple regression procedure. To be included in the resulting regression model, a variable had to be significant at the  $p < .15$  level.

The primary purpose of the current study was to determine if controllers who differed in their reported awareness of the error also differed in the types of errors they made. We began by trying to predict the severity of the operational error from 1) other characteristics of the error, 2) personnel variables, 3) situation variables, and 4) causal factors. Finally, we explored whether aware controllers and unaware controllers differed in the type of psychological process that went awry during the error. If reported awareness is an unimportant factor, then we should see little difference between aware and unaware controllers.

### Predicting Error Severity

#### Variables characterizing operational errors

To avoid an error in en route airspace, aircraft must be 5 NMnm apart horizontally or 1000 feet (2000 over FL 290) apart vertically. Violations of these minima yield errors of different severity. *Severity*, as used here and elsewhere (Rodgers & Nye, 1993), is the result of a configurational rule for combining violations of vertical separation and horizontal separation. Scores range from 0 to 20.

For example, a major error is given 20 points and occurs when aircraft are separated by less than 500 ft vertically and less than .5 miles horizontally. Moderate errors have severity scores ranging from 14 to 19. Minor errors have a severity scores of 13 or less. In our 1993 dataset, there were no major errors.

Virtually all of the errors reported were exclusively human (99%), that is did not involve equipment or procedures, involved only one facility (96%), and were not related to procedural deficiencies (98%). The flights involved in the errors we examined were predominantly under radar control (99%) and typically one flight was descending (83%) either through an altitude occupied by a level flight (42%) or through airspace used by a climbing flight (33%).

Sixty-two percent of the controllers were not aware that an operational error was developing, whereas 38% were aware of the error but were unable to rectify it in time. Typically, the errors were ultimately recognized by means of the conflict alert (46%) or were self-identified (40%).

Predicting severity from the set of error cohort variables using stepwise multiple regression, yielded an equation of three predictor variables accounting for 6.8% of the variance in severity: Controllers who were aware that an error was developing made significantly less severe errors (8.93, range 3 - 17) than those who were unaware (10.66, range 3 - 19). The other two variables of the model, each accounted for less than 1% of the variance. Procedural deficiency, an indicator of deficiencies in established procedures, was more likely to be implicated for the more severe errors. Finally, although few incidents involved nonradar flights, those that did yielded significantly more severe errors. Operational errors involving the four nonradar flights (Mean severity = 13.25) were more severe than were errors involving radar flights ( $M = 9.96$ ).

Because these variables occurred simultaneously with the error, or could be identified only after the error occurred, they are not of great use in identifying variables that could be changed or modified. However, as statistical predictors, they could be used in considering subsequent analyses. For example, in this report, we take advantage of the large difference between controllers who were aware of the impending problem and those who were not.

Because we were particularly interested in factors that could be indicators of poorer SA/situation awareness, in the following multiple regression analyses, we considered separately the 150 reports for which the controller was aware, and the 238 reports for which the controller was unaware. We did this, in part, because it was impossible to predict *awareness* from variables about the personnel or about the situation. No regression model of personnel variables emerged that allowed us to predict who would be aware of an impending error and who would not. For a situation based model, the type of control procedures involved (e.g., radar, nonradar) and number of persons involved were implicated, but together they accounted for only 1.6% of the variance in awareness. Thus, given the large impact of awareness on severity, and given the difficulty of predicting awareness from other variables existing before the error, we instead analyzed controllers who were aware of the error separately from those who were not. For ease of exposition, we will often call the former group *aware controllers* and the latter group *unaware controllers*.

#### Personnel variables

Controllers who made operational errors in 1993 were, on the average (median), full performance level controllers (FPLs) for 41 months (0 to 305) who had last been (re)certified 30 months earlier (0 to 293). About half had medical waivers or restrictions (47%). Typically, they had just had a regular day off (RDO) the day before (30%) or two days before (26%) and had returned from a break from controlling traffic 41 minutes before (1 to 395).

Stepwise multiple regressions were used to predict the severity of the error given the personnel variables. For those controllers aware of the impending error, length of time as an FPL was the only contributor to the model, accounting for only 2.8%. The model for unaware controllers accounted for an equally unimpressive 2.4%. Like aware controllers, the longer an unaware controller was an FPL the less severe the error; unlike aware controllers, time since certification was a factor for unaware controllers: The more recently the controller was (re)certified, the less severe the error. This latter finding is somewhat surprising, and perhaps suggests that controllers

who were recently required to go through the recertification process might be somewhat more attentive or more careful. A controller can go through the certification process for several reasons, including switching facilities or areas of specialization. Regardless, the ability to predict the severity of operational errors from personnel variables is disappointingly weak.

#### Situation variables

Variables characterizing the situation in which the operational error occurred did not suggest that these errors took place in obviously unusual situations. Traffic complexity was average (3, on a scale ranging from 1 to 5) with approximately 8 aircraft (2 to 30) under radar procedures (99%) in positive control airspace (54%). Sectors were usually not combined (74%), nor were positions (56%), no training was in progress (85%), no special procedures were in effect (86%), and national procedures were being followed (99%). More errors (48%) tended to occur near the end of the calendar year (July through November).

Situation variables did not prove good predictors of error severity. For aware controllers, a simple model including only the presence of special procedures emerged from the analysis, accounting for 3.4% of the variance. Errors with special procedures were worse. For unaware controllers, no model emerged from the 11 situation factors that were used to predict severity.

#### Causal factors

The investigators who filed the operational error report attempted to discern the causal factor(s) responsible for the error. Table 1 lists the causal factors with the percentage of times they were implicated in an error. The most frequently identified causal factor was other inappropriate use of display data (IUDD), which was implicated 44% of the time. This other factor was often something rather uninformative, such as the controller failed to recognize traffic; failed to observe aircraft was 5NM east of J48, and so on. The second most frequent cause, inappropriate use of Mode-C, was implicated 10% of the time.

We began by comparing the proportion of times a factor was implicated for aware and unaware controllers. Only two causal factors were reliably ( $p < .05$ ) different: Aware controllers were more likely to have made an other IUDD error ( $\chi^2(1) = 4.96$ ) than were unaware controllers, although they also made this type of error more than any other. The unaware controllers were more likely than were aware controllers to commit an error involved with readback of altitude ( $\chi^2(1) = 6.60$ ).

Table 1. Percentage of uUnaware and aAware controllers reports implicating each causal factor.

Causal Factors	% Unaware	% Aware
None	2.1	4.7
Computer Entry		
Incorrect input	2.1	0.7
Incorrect update	3.3	0.7
Premature termination of data	1.7	0.0
Other (Explain)	5.0	1.3
Flight Progress Strip		
Not updated	4.2	1.3
Interpreted incorrectly	1.3	6.0

Posted incorrectly	0.4	0.0
Updated incorrectly	1.7	1.3
Premature removal	0.4	0.0
Other (Explain)	0.8	1.3
Radar Display/Misidentification		
Failure to reidentify aircraft when the accepted target identity becomes questionable	0.4	0.0
Overlapping data blocks	2.1	2.7
Other (Explain)	10.0	6.0
Radar Display/Inappropriate Use of Displayed Data		
Mode C	6.3	12.7
BAFE	0.4	2.7
Conflict alert	2.9	1.3
Other (Explain)	39.2	49.3
Communications Error		
Phraseology	4.2	2.7
	7.1	3.3

WISCONSIN  
AIRCRAFT

4.7  
4.7

12.5	3.3
4.6	1.3
1.1	0.0
3	0.0

0.8	0.7
0.8	0.0
0.4	0.0
1.3	0.0
1.3	0.7

between Aircraft #1 and Aircraft #2; Employee assumed Aircraft #2 would stay on an easterly heading after deviating around weather). Uninformative narratives were those that provided no insight into the underlying cognitive process and were excluded from analyses; forty-one percent of the aware reports and 24% of the unaware reports were excluded. A chi-square analysis of awareness (i.e., unaware vs. aware) conducted for each of the four psychological classifications revealed a significant dependent relationship. In comparison to aware controllers, unaware controllers made significantly more perceptual, ( $\chi^2(13) = 410.181, p < .05$  and memory errors ( $\chi^2(1) = 6.00$ ). In contrast, aware controllers committed significantly more thinking errors,  $\chi^2(1) = 9.26$ . The results are presented in Figure 1. Interestingly, attentional factors were implicated equally often in aware and unaware controllers operational errors. However, memory factors were implicated considerably more often for unaware controllers compared with aware controllers, whereas thinking factors were implicated much more often for aware than unaware controllers. Again, the data seem clearer for aware controllers than for unaware controllers. Aware controllers make a considerable number (56%) of thinking errors, errors of assumption, judgment, and decision making. However, the types of errors committed by unaware controllers are more evenly distributed across the defined psychological categories. Unaware controllers make relatively more memory errors than aware controllers, but in fact the frequency of memory errors is no higher than perceptual or thinking errors. Unaware controllers seem to make all types of errors.

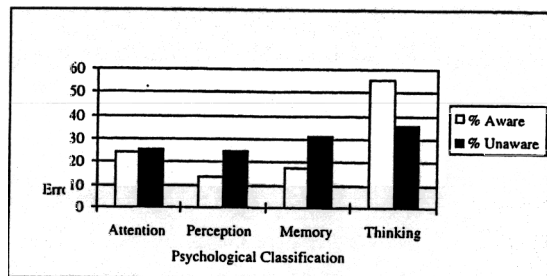


Figure 1. Psychological classification by awareness

## Discussion

How might aware controllers differ from unaware controllers? The current work allows some speculation. One possibility is that aware controllers had reasonably good SA, but because of a misjudgment or false assumption, they inadvertently performed one action, when they should have performed another. Given that the only situation variable for aware controllers was the presence of special procedures, it may be that aware controllers incorrectly defaulted to more standard procedures, rather than implementing the special procedures. In a schema view, this would mean that the controller relied on default values in their schema, or default settings in their mental model. This would lead to errors of interpretation and judgment. Although one could speculate in this

way, this type of speculation is premature, due in large part to the lack of any understanding of situation factors influencing the unaware controller.

Some understanding of the difference between aware and unaware controllers comes may emerge from the analysis of the narratives produced by the quality assurance specialists. A gross differentiation along cognitive dimensions suggested that aware controllers made errors in thinking (i.e., judgments, decision making, and assumptions), whereas the errors made by unaware controllers tended to involve memory. Although aware and unaware controllers were different in the narrative analysis, and although aware controllers tended to make thinking errors, it is again less clear what to make of the data for unaware controllers. It is possible that unaware controllers made more memory and perceptual errors, but it is also possible that unaware controllers made errors across the board, in roughly equal frequencies. Overall, as with some of the other analyses (e.g., situation-based stepwise regression), the characteristic error made by the unaware controller is not readily apparent from the retrospective operational error report. It seems clear that any fine-grained understanding of situation awareness should rely upon on-line measures, rather than retrospective reports.

## References

- Redding, R. E. (1992). Analysis of operational errors and workload in air traffic control. In *Proceedings of the Human Factors Society 36th Annual Meeting*, 2, 1321-1325. Santa Monica, CA: Human Factors Society.
- Rodgers, M. D., & Nye, L. G. (1993). Factors associated with the severity of operational errors at air route traffic control centers. In Rodgers, M. D. (Ed.), *An examination of the operational error database for air route traffic control centers* (DOT/FAA/AM-93/22, pp. 11-25). Washington, D. C.: Office of Aviation Medicine.
- Schroeder, D. J., & Nye, L. G. (1993). *An examination of the workload conditions associated with operational errors/deviations at air route traffic control centers* (DOT/FAA/AM-93/22, pp. 1-9). Washington, D. C.: Office of Aviation Medicine.

## Acknowledgments

This research was supported by contract # DTFA-02-94-T-80261 from the Federal Aviation Administration to Francis T. Durso. Address correspondence to Frank Durso, Department of Psychology, University of Oklahoma, Norman, OK 73019-0535 or e-mail: fdurso@uoknor.edu.

## Expertise and Chess: A Pilot Study Comparing Situation Awareness Methodologies

Francis T. Durso<sup>1</sup>, Todd R. Truitt<sup>1</sup>, Carla A. Hackworth<sup>1</sup>,  
Jerry M. Crutchfield<sup>1</sup>, Danko Nikolic<sup>1</sup>, Peter M. Moertl<sup>1</sup>,  
Daryl Ohrt<sup>1</sup>, and Carol A. Manning<sup>2</sup>

<sup>1</sup> University of Oklahoma

<sup>2</sup> F.A.A. Civil Aeromedical Institute

Situation awareness (SA) has received a considerable amount of attention in the recent literature. However, no agreed upon definition or methodology for the measurement of this phenomenon currently exists. Many definitions have been proposed and have provided perspectives varied in scope. For example, the definitions provided by Endsley (1988) and Mogford (1994), although not incompatible, present different viewpoints regarding SA. Despite the lack of a common definition, our experiment focused on finding a sensitive procedure that would best differentiate among levels of expertise or skill and hence, SA.

A number of different methodologies have been explored in an effort to understand how operators develop and maintain a "picture" of the situation in which they are involved. Methodologies previously used have included verbal protocol analysis (e.g., Ericsson & Simon, 1994; Ohnemus & Biers, 1993; Sullivan & Blackman, 1991), retrospective event recall (e.g., de Groot, 1965; Kibbe, 1988), concurrent memory probes such as the *Situation Awareness Global Assessment Technique* (SAGAT; Endsley, 1988), and physiological measures such as eye movements (e.g., Moray & Rotenberg, 1989; Stein, 1989; Wierwille & Eggemeier, 1993). Beginning with the assumption that experts have better SA than novices, we compared five very different procedures in one study: verbal protocols, eye movements, on-line queries with the situation present, on-line queries with the situation removed (as in SAGAT), and post-hoc recollection.

We chose to look at chess expertise for several reasons. First, chess has been correctly called the drosophila of cognitive psychology (Charness, 1989), and its long history of study should serve us well in understanding SA. Second, it seems to us that, perhaps more than most activities, differences in chess expertise are differences in SA. The entire game involves assessing the relationship between existing pieces and predicting impending moves. For example, input and output processes are simple and are unlikely to distort the internal model of the situation. Third, chess players are ranked by the United States Chess Federation (USCF) and the differences among rankings are well understood. Finally, our ultimate interest is in understanding SA in air traffic controllers, and chess provides a God's-eye perspective of a number of different entities that make it a nice, albeit limited, laboratory analog of air traffic control.

### General Methodology

All participants monitored four chess games. We asked participants to monitor, rather than play, the games to allow control over the entire game. All players experienced exactly the same game

regardless of their skill level, allowing us, for example, to insert queries at identical points for each player. To ensure involvement, the participants were asked to monitor the game for imminent material losses. This is a clear component of chess and allowed us to engage the participants even though they were simply monitoring an existing game.

Table 1. Overview of experimental procedures.

	Game 1	Game 2	Game 3	Game 4
Opening	Queen's Gambit (Slav defense/Main line)	Queen's Indian (Main line)	Caro-Kann (Capablanca/Main line)	Sicilian (Alapin)
Outcome	White mates in 60	Black mates in 74	White mates in 52	Black mates in 61
Number of captures	8	7	12	8
Monitor material loss	Yes	Yes	Yes	Yes
Method tested	Verbal protocols	Eye movements	Situation-present queries	Situation-absent queries (SAGAT like)
Post-hoc recall	Yes	Yes	Yes	Yes

In each game, the participant sat 42" from a projected image of a chess board that subtended a visual angle of 40°, with each square subtending about 5°. At the beginning of each ply, a piece blinked twice, moved, and then flashed twice again. The position remained for 15 sec. The game was stopped after 80 plies (40 moves).

A chess expert (USCF 2100; Expert), an intermediate (USCF 1607; Class B), and a novice (USCF 1374; Class D) monitored four high-level games generated by a commercial computer program, *Chessmaster 4000*. Characteristics of each game, and what methodological procedures were employed, appear in Table 1. The games were generated by setting both white and black on *Chessmaster's* Chessmaster-level and having the computer play itself. Games were randomly assigned to order of presentation. In all games, moves of both white and black were made while the participant monitored the game to determine when a piece was about to be captured. Across games there were 7 to 12 plies or sequences of plies during which pieces were captured.

In this pilot study, the use of only three participants makes the statistical discovery of the most sensitive procedure problematic. In part we relied on visual inspection of the data, looking for clear differences across levels of expertise. In addition, we conducted item analyses using materials (e.g., queries) rather than participants as the random variate. This allows us to generalize to other probes or situations for these participants, but does not allow us, for example, to discriminate between effects caused by the particular participant and effects caused by his level of skill. All tests were conducted at an alpha level of .05.

## Anticipating Material Loss

As the cover task, the participant used a joystick to register his judgment about the imminent loss of material. If he believed that a piece would be taken "in the near future", he was to pull the joystick back; if he believed that a piece would not be taken in the near future, he pushed the stick forward. "Near future" was intentionally left vague so as not to influence the distance in the future that the participant normally considered. Confidence was indicated by the extent to which the stick

was pushed or pulled. If the participant did not have any feeling about the upcoming state of affairs, he was to rest the stick in the middle, neutral position.

We began analysis by determining the plies during which a take occurred. To control for strategies (such as always anticipate loss of material) we chose, for each game, a comparable number of plies in which a take did not occur. Thus, if a person did invariably pull the joystick back, he would do well anticipating loss of material, but poorly anticipating when no material loss would occur. Performance of a theoretical perfect player who moved optimally from take to no-take and back was calculated. For each critical ply in the game, we computed the point prior to that occurrence when the participant changed his judgment from "no take" to "take," or from "take" to "no take," and compared it to the theoretical perfect participant. Scores could range from 0 sec, if the ply was never correctly detected, to the theoretical optimum.

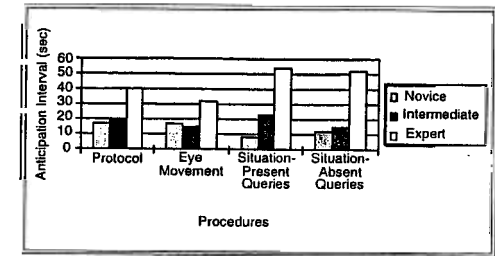


Figure 1. Projecting material loss

Not surprisingly, the expert anticipated material losses (47 sec; 4 plies) sooner than did the class B player (19 sec; 2 plies) or the class D player (13 sec; 1 ply). Results are shown in Figure 1. Regardless of the concurrent procedure, the expert was easily distinguishable from the two nonexperts. A repeated measures ANOVA revealed a significant main effect of skill,  $F(2, 132) = 14.82$ . Although there was no significant effect of method nor any interaction of method with skill, the figure suggests that only when the participants were in the Situation-present Queries condition were all three levels of expertise distinguishable. This may suggest that the Situation-present Queries condition is most likely to reflect individual differences in SA.

There has been some concern that interrupting the participant could, in itself, interfere with SA (e.g., Sarter & Woods, 1991). However, in the current study, there was little indication for the novice or the intermediate that the particular procedure had any contaminating effect on SA. While verbal protocols have been criticized as obtrusive and unreliable (Nisbett & Wilson, 1977), recording eye movements is touted as a relatively unobtrusive measure of recording behavioral data. However, in our data there was no effect of methodology on anticipating material loss, and what little effect may have been present argues for the use of query techniques.

## Verbal Protocols

Verbal protocols were gathered during the first game. Participants were told to "... talk out loud. Try and verbalize your thoughts about the chess game that you are watching. For example, talk about pieces that are about to be taken, tactics or strategies that you notice are being used, or just general comments about the game such as which side currently has an advantage. We do not expect you to comment about anything in particular, just tell us basically what you are thinking about regarding the game...."

Protocols were transcribed, segmented into plies, and then categorized by the experimenters. Comments relevant to the game were coded into four categories: Predictions, Assessments, Identifications, and Other. Predictions were sentences that described possible future moves (e.g., "The rook will go to H3 and attack the queen."). Assessments were sentences that characterized the ongoing flow of the game without making any predictions (e.g., "White needs to avoid exchanges because he has the developmental advantage."). Identifications were utterances that identified specific moves or tactics (e.g., "The knight is pinned; This is a Queen's gambit."). Comments that did not fit these categories were rare and were excluded from further analyses.

Overall, the intermediate participant made the most utterances ( $N = 184$ ), the expert the least ( $N = 94$ ), with our novice ( $N = 106$ ) falling between these two. This pattern of overall utterances was consistent with the intermediate effect (Grant & Marsden, 1988; Schmidt & Boshuizen, 1993). When the utterances are classified as in Figure 2, it becomes apparent that the participants uttered different types of comments depending on their level of expertise,  $\chi^2(4) = 9.49$ . Our expert produced mostly predictions (64%), more than either the intermediate or the novice,  $\chi^2(2) = 12.46$ . The intermediate, on the other hand, produced the largest number of statements assessing the situation,  $\chi^2(2) = 76.03$ . His corpus was 62% assessment, with only 27% predictions and 10% identifications. Finally, the novice's protocol consisted of 50% assessment, 25% prediction, and 19% identification, a profile too similar to the intermediate's to allow recommendation of this procedure, resulting in a nonsignificant  $\chi^2(2) = 4.0$ , *NS*.

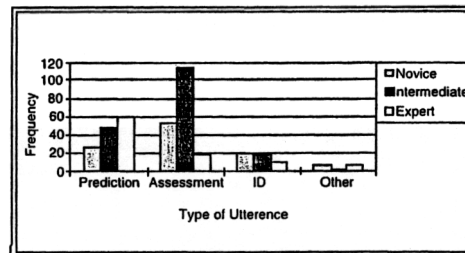


Figure 2: Protocol analysis

## Eye Movements

During game two, eye movements were recorded using an Applied Science Laboratories (ASL) series 4000 eye tracker. The eye tracker utilized a magnetic headtracking apparatus in order to compensate for any head movements made by the participant. Eye movement data were recorded in two, 10 minute sessions. The first session included plies 1-40; the second session included plies 41-80. A 10 minute break was allowed between sessions to prevent participants from experiencing any discomfort resulting from the headband of the eye tracker. Participants were instructed to watch the game while using the joystick as in the previous game. Fixations and saccades were recorded. To qualify as a fixation, the eye had to remain looking within .5° visual angle for at least 100 msec.

Unfortunately, in all of our analyses, differences among skill levels were quite small. For example, the largest difference in the fixation rate was 0.4 fix/sec, and the difference in average saccade distance was 1°. A more promising finding may be that the novice player had longer fixation durations ( $M = 283$  msec) than did the expert ( $M = 233$  msec) or intermediate player ( $M = 198$  msec), but even here it was difficult to clearly classify the players. We looked at a myriad of other measures (e.g., fixation duration on critical pieces, area of the board covered) with little evidence that eye movements would be a consistent predictor of expertise. It certainly would not be an easy one to ascertain. Unfortunately, a clear picture of SA, as reflected in eye movement differences, was not apparent, although some interesting trends did emerge. Obviously, this is not to say that such differences do not exist, merely that we could not find them.

## Situation-present Queries

During the third game, the participant responded to questions about current and future chess positions. On some trials during the chess game, a tone sounded and a question was presented auditorily while the chess board and pieces remained in view. All participants were asked the same questions on the same plies. Eighteen questions were asked; six from each of three categories: 1) Perceptual, 2) Present Conceptual, 3) Future Conceptual. In this version of the on-line query methodology, the situation remains present while the participant responds. Clearly, the proportion of correct responses should be quite high, given that the participant can determine the correct answer from information still being displayed. Thus, the primary dependent variable was response time.

We looked at this variation of more traditional on-line querying techniques for a number of reasons. First, unlike looking at mistakes in SA, response time allows us to investigate successful SA, rather than inferring characteristics of SA from its failures. Second, our ultimate interest is in exploring SA among air traffic controllers, where any technique that removes them from the radar screen is likely to be disruptive and viewed with suspicion. An auditory query allows the situation to continue and uses an input mode that can be fit into the controller's existing work scheme (e.g., by querying over a telephone line).

The participant answered orally while the experimenter recorded his response. The screen in front of the participant went blank and the response recorded by the experimenter was then displayed, allowing the participant to confirm or change the experimenter's input. A question was asked during 18 randomly selected plies during the game. Participants were queried about: 1) Perceptual characteristics (Where is the white queen?; What piece is adjacent to the black rook?; 2) Present conceptual relations (What piece is the white bishop attacking?; What piece is defending the white knight?; 3) Future relations (What piece can white move to pin black's rook?; What piece can black move to prevent a back rank mate?)

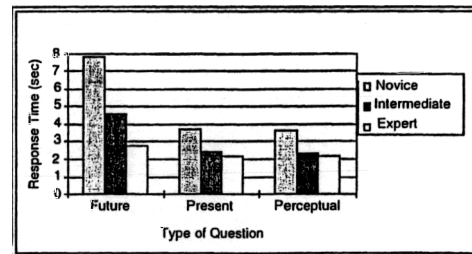


Figure 3. Situation-present queries

As expected, accuracy was quite high and did not distinguish among levels of expertise. Response time, on the other hand, appears to be a good index of expertise, and presumably, SA,  $F(2, 24) = 13.12$ . Results are shown in Figure 3. Overall, the expert responded faster than the intermediate, who responded faster than the novice. Latency also varied as a function of question type,  $F(2, 12) = 4.78$ , with future questions taking longer to answer than did the other types. Skill also interacted with question type,  $F(4, 24) = 2.32$ ,  $p < .10$ . Differences as a function of skill were present most clearly in questions about future events,  $F(2, 8) = 7.20$  and perceptual events,  $F(2, 10) = 5.28$ .

### Situation-absent Queries

More typically, on-line queries freeze the simulation of interest, remove information, and ask the participant a question or series of questions (i.e., Endsley's (1988) SAGAT). In the fourth game, the pieces were removed from the board as a tone sounded, and a question was presented visually to the right of the now empty board. Participants read and answered the question, the screen went blank, the experimenter typed the response to allow the participant to verify the entry, and then the board reappeared as it was when the question sequence was initiated. As before, 18 questions were asked.

The data from situation-absent procedures are the number of questions responded to correctly. Those data appear in Figure 4. It is worth noting that the figure presents results similar to those found in the verbal protocol: an expert advantage for predictions (future) and an intermediate advantage for assessments of the (current) position. This methodology showed a marginal skill effect,  $F(2, 30) = 2.94$ ,  $p < .10$ . Questions about the future distinguished most clearly among the three skill levels,  $F(2, 10) = 3.18$ ,  $p < .10$ . Skill differences were not reliable for the perceptual or present questions, despite the graph's suggestion of an intermediate superiority for present-queries.

Given the success we had with the response time analysis in the situation-present procedure, we looked at the response times associated with the correct responses in the situation-absent procedure. Unlike percent correct, visual inspection of the response time data revealed consistent differences among skill levels (although not as clear as the situation-present condition). Unfortunately, the error rate prevented any meaningful analysis of correct latency. Nevertheless, this may suggest that the differences between the situation-present and situation-absent procedures

may be relatively unimportant, provided that response time is used to assess differences among levels of expertise.

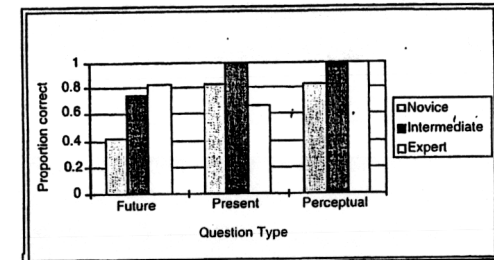


Figure 4. Situation-absent queries

### Memory

The final analysis involved consideration of the recall protocols participants gave after the game. Although recall occurred after each game, we considered only the first game here, since it would be uncontaminated by other recalls. The participants differed dramatically in the length of their recall protocols, with the expert saying very little and the intermediate saying quite a bit. However, the expert's recall consisted primarily of general abstractions that characterized large segments of the game. The novice's recall consisted of several move by move recollections of the game. The intermediate's recall fell between these two. Unfortunately, the succinctness of the expert's recall made it difficult to analyze the protocols beyond this overall classification. However, it does suggest that experts tend to convey labels of encapsulated information (Schmidt & Boshuizen, 1993).

### Discussion

We investigated five procedures. Of those, eye movements seemed to be the most complicated and yielded the fewest insights. Analysis of memory protocols and on-line protocols were also problematic. The two query procedures seemed to supply some useful information. Across the procedures, we found considerable evidence that questions about the future are most likely to discriminate among all three levels of expertise. Number of predictions in the verbal protocols, proportion correct for future-oriented queries (situation-absent), and response time for future-oriented queries (situation-present) all suggest that researchers interested in distinguishing levels of SA would do well to focus their efforts on future events. Even our cover task, which required participants to predict material loss, showed clear expertise effects. Information about the current state of affairs, as measured by assessment utterances in verbal protocols and proportion correct in



situation-absent queries, did not discriminate across skill levels. This was due to the finding of an intermediate effect where the intermediate chess player outperformed both the expert and the novice. The problem in assessing SA with these types of information is that the novice and the expert are more similar than their ranking suggest they should be.

The current work also suggests that the best procedure for measuring SA is to ask questions of the participant, but to rely on the response time for correct responses, rather than on the proportion of correct responses. In the situation-present procedure, response time (especially for future events) was a good discriminator of skill level. Even in the situation-absent procedure, response time may prove to be a fairly good discriminator. The query procedures also did not seem to fetter the participants' ability to monitor the task, as evidenced by their continued ability to project material loss throughout the procedure. This suggests that interrupting participants with a query does not seriously disrupt the task. Further, there seems to be no reason to remove the situation in order to assess SA. Of course, proportion correct will be near ceiling since the participant can simply reexamine the display; however, response time will capture SA differences. In fact, it will capture it more clearly than removing the situation. We believe this finding to be quite encouraging for efforts, such as ours, that hope to be applicable to air traffic controllers. We believe, in such an environment, an important part of SA is knowing where to find the information when you need it; in air traffic control, not all information needs to be memorized, nor should it be. With the situation present, the controller can continue to control traffic while SA queries from other facilities are occasionally presented over one of the controller's telephone lines. Attempts to extrapolate these results to other arenas should take into consideration the small number of participants used in this study, the differences between task domains, and the differences due to monitoring versus active control. Nevertheless, the current work suggests that the best way to assess SA is to ask questions about impending events while the situation remains available to the operator and to time how long it takes the operator to make the correct response.

## References

- Charness, N. (1989). Expertise in chess and bridge. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 183-208). Hillsdale, NJ: Erlbaum.
- de Groot, A. (1965). *Thought and choice in chess*. Paris, France: Mouton.
- Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, (pp. 97-101). Santa Monica, CA: Human Factors Society.
- Ericsson, K., & Simon, H. (1984). *Protocol analysis - Verbal reports as data*. Cambridge, MA: MIT Press.
- Grant, J., & Marsden, P. (1988). Primary knowledge, medical education and consultant expertise. *Medical Education*, 22, 746-753.
- Kibbe, M. P. (1988). Information transfer from intelligent EW displays. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, (pp. 107-110). Santa Monica, CA: Human Factors Society.
- Mogford, R. H. (1994). Mental models and situation awareness in air traffic control. In R. D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems*, (pp. 199-207). Embury-Riddle Aeronautical Press.
- Moray, N., & Rotenberg, I. (1989). Fault management in process control: Eye movements and action. Special Issue: Current methods in cognitive ergonomics. *Ergonomics*, 32, 1319-1342.
- Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.

- Ohnemus, K., & Biers, D. (1993). Retrospective versus concurrent thinking-out-loud in usability testing. In *Proceedings of the 37th Annual Meeting of the Human Factors Society*, (pp. 1127-1131). Santa Monica, CA: Human Factors and Ergonomics Society.
- Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, 1, 45-57.
- Schmidt, H. G., & Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory & Cognition*, 21, 338-351.
- Stein, E. S. (1989). *Air traffic controller visual search* (DOT/FAA/CT-TN89/9). DOT/FAA Technical Center: Atlantic City International Airport, New Jersey.
- Sullivan, C., & Blackman, H. S. (1991). Insights into pilot situation awareness using verbal protocol analysis. In *Proceedings of the Human Factors Society 35th Annual Meeting*, (pp. 57-61). Santa Monica, CA: Human Factors Society.
- Wierwille, W., & Eggemeier, F. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35, 263-282.

This research was supported by contract # DTFA-02-94-T-80261 from the Federal Aviation Administration to Francis T. Durso. Address correspondence to Frank Durso, Department of Psychology, University of Oklahoma, Norman, OK 73019-0535 or e-mail: fdurso@uoknor.edu.